Deriving a lexicon for a precision grammar from language documentation resources: a case study of Chintang

Bender, Emily M ; Schikowski, Robert ; Bickel, Balthasar

Abstract: Language documentation projects typically invest a lot of effort in creating digitized lexical resources, which are used in the creation of dictionaries and in the glossing of collected texts. We present and evaluate a methodology for repurposing such a lexical resource developed for Chintang (ISO639-3: ctn), a language of Nepal, for use with a precision implemented grammar developed in the DELPH-IN formalism. The target lexicon, when combined with a set of morphological rules, achieves 57% type-level coverage and 50% token-level coverage of held-out texts, while maintaining a feature-level accuracy F-measure of 70%. As lexicon development is typically one of the most expensive aspects of creating a precision grammar, this represents a significant savings of effort.

Originally published at:
Bender, Emily M; Schikowski, Robert; Bickel, Balthasar (2012). Deriving a lexicon for a precision grammar from language documentation resources: a case study of Chintang. In: Kay, M; Boitet, C. Proceedings of the 24th International Conference on Computational Linguistics (COLING). Mumbai: Association for Computational Linguistics, 247-262.

Deriving a Lexicon for a Precision Grammar from Language Documentation Resources: A Case Study of Chintang

Emily M. BENDER[1] *Robert SCHIKOWSKI*[2]
Balthasar BICKEL[2]
(1) Department of Linguistics, University of Washington
(2) Seminar für Allgemeine Sprachwissenschaft, Universität Zürich
ebender@uw.edu, robert.schikowski@uzh.ch, balthasar.bickel@uzh.ch

ABSTRACT
Language documentation projects typically invest a lot of effort in creating digitized lexical resources, which are used in the creation of dictionaries and in the glossing of collected texts. We present and evaluate a methodology for repurposing such a lexical resource developed for Chintang (ISO639-3: ctn), a language of Nepal, for use with a precision implemented grammar developed in the DELPH-IN formalism. The target lexicon, when combined with a set of morphological rules, achieves 57% type-level coverage and 50% token-level coverage of held-out texts, while maintaining a feature-level accuracy F-measure of 70%. As lexicon development is typically one of the most expensive aspects of creating a precision grammar, this represents a significant savings of effort.

TITLE AND ABSTRACT IN GERMAN

Ableitung des Lexikons für eine Präzisionsgrammatik aus dokumentationslinguistischen Ressourcen anhand einer Fallstudie zum Chintang

Typische Sprachdokumentationsprojekte investieren viel Zeit in den Aufbau digitaler lexikalischer Ressourcen, die für die Erstellung von Wörterbüchern und für die Glossierung von Korpustexten genutzt werden können. Dieser Vortrag stellt eine alternative Verwendung eines elektronischen Wörterbuchs vor, das für das Chintang (ISO639-3:ctn), eine bedrohte Sprache Nepals, entwickelt wurde. Die Kombination dieses Wörterbuchs mit einer nach dem DELPH-IN-Formalismus entwickelten Präzisionsgrammatik in Form morphologischer Regeln kann erste Texte auf der Type-Ebene zu 57% und auf der Token-Ebene zu 50% abdecken, wobei auf der Merkmalsebene ein F-Maß von 70% gewahrt wird. Da der Aufbau lexikalischer Ressourcen zu den zeitintensivsten Komponenten der Entwicklung einer Präzisionsgrammatik gehört, bringt diese Methode eine signifikante Zeitersparnis mit sich.

KEYWORDS: lexical acquisition, grammar engineering, endangered languages, low-resource languages, language documentation.

KEYWORDS IN GERMAN: Lexikonerstellung, Grammar Engineering, bedrohte Sprachen, Sprachen mit geringen Ressourcen, Sprachdokumentation.

Proceedings of COLING 2012: Technical Papers, pages 247–262,
COLING 2012, Mumbai, December 2012.

1 Introduction

Endangered languages represent an especially urgent type of low-resource languages: Not only do they generally lack computational resources, but they also unfortunately have the property that the window in which to create such resources is small and closing. Thus to the extent that any computational resources are created, they are particularly valuable, and if they can be repurposed for applications beyond their original target, this extends their value further.

This paper describes a case study in the repurposing of a lexical resource for an endangered language, Chintang. The original lexical resource is a Toolbox[1] lexicon developed to assist in the glossing of collected texts and in the development of a conventional dictionary. We present a methodology for automatically translating this lexicon into one which can be used as part of a precision grammar for the language, suitable for both parsing and generation and eventually in applications including machine translation. The grammar is a starter grammar generated by the LinGO Grammar Matrix grammar customization system (Bender et al., 2010), and includes an initial implementation of Chintang verbal morphology. As a starter grammar, it still has very limited coverage. However, even broad coverage grammars rely heavily on the size and quality of their lexicons. Our focus here is therefore on the extent to which the existing resources for Chintang can be used to bootstrap a lexicon for this implemented grammar.

This paper is structured as follows: In Section 2, we provide background on the Chintang language, the project which is documenting it, and the relevance of precision grammars for language documentation. Section 3 describes our methodology for creating the Grammar Matrix-compatible lexicon on the basis of the Toolbox lexicon and the implementation of the morphology. We evaluate the lexical coverage of the resulting grammar over held-out texts in Section 4. Section 5 situates this work with respect to related initiatives.

2 Background

2.1 Chintang Language

Chintang (ISO639-3: ctn) is a Kiranti language spoken in the hills of Eastern Nepal. The next bigger city close to the village of Chintang is Dhankuta, which is a six hours footwalk away. The local economy is centered around agriculture, both for subsistence and for trade.

The Chintang language belongs to the large Sino-Tibetan (Tibeto-Burman) family. The exact position of the Kiranti languages within this family is unclear (Ebert 2003). Within Kiranti, Chintang belongs to the Eastern branch, which is characterized by the development of the preglottalized stops reconstructed for Proto-Kiranti by Michailovsky (1994) to aspirated stops.

The number of speakers of Chintang is generally estimated to be around 5000 (e.g. Bickel et al. 2007), and this is in accordance with the speakers' own estimations. There are no reliable official data. Most speakers are bilingual, speaking Nepali, the national language of Nepal, as the second language. Many in addition speak Bantawa, a big neighboring Kiranti language. Chintang is still being learned by children (Stoll et al., 2012), but language knowledge and transmission are clearly on the decline, especially in the more easily accessible parts of the village.

[1] http://www.sil.org/computing/toolbox/index.htm

2.2 Research projects and resources

Research on Chintang started with the Chintang and Puma Documentation Project (CPDP), which ran from 2004 to 2009. Since 2009, research has been continued by a group of several collaborative projects together referred to as the Chintang Language Research Program (CLRP).[2] Several native speakers of Chintang and Nepali were employed during the projects to make recordings, conduct interviews, and to transcribe and translate recordings. Presently there is an office at the Tribhuvan University, Kathmandu, where five transcribers and translators are constantly working on the corpus, but no new recordings are being made.

The corpus comprises about 280 hours of video recordings, the majority of which have been transcribed by now (250 hours, containing 1,130,000 words; Bickel et al., 2009ff). Transcribed sessions are first translated from Chintang to Nepali and then from Nepali to English. The English translation is an important aid for the final step of glossing, which is done by student assistants studying linguistics. So far approximately 620,000 words have been glossed. Additional annotations are added to parts of the corpus depending on the needs of individual projects. Examples include the annotation of pointing gestures or of referential properties such as identifiability.

The compilation of the corpus is tightly coupled with the Chintang dictionary, which presently has about 9,000 words. The electronic version was created along with the corpus, so all words in the corpus are in the dictionary. Some systematic elicitation work to cover semantic fields that do not frequently come up in everyday conversation was carried out in 2010, and a printed version for the speaker community was published in 2011 (Rāī et al. 2011). The electronic dictionary keeps growing as more and more words are glossed. New words collected by the glossers are integrated into the main dictionary twice a year.

Both the corpus and the dictionary are in Toolbox format. In Toolbox, files are divided into structurally similar records (utterances in the case of the corpus, entries in the case of the dictionary). Each record consists of several lines where each line starts with a so-called field marker indicating the type of information (e.g. phonological words, morphemes, morpheme glosses) followed by content (see Section 3 below for details). It is possible to align the tiers thus defined, enabling composite searches (e.g. "find all morphemes of the shape *cekt* which have been glossed as 'speak'"). Since a major revision of the dictionary in 2010, all dictionary entries have IDs, which are inserted and aligned with morphemes upon glossing. This makes it possible to automatically look up detailed information for each corpus morpheme in the dictionary.

The entries in the dictionary include stem forms, alternate forms, glosses in English and Nepali, as well as some grammatical information. In particular, the dictionary lists coarse-grained part of speech (drawn from a set of 30 tags) as well as fairly detailed information about the syntactic and semantic valence of verbs, that is, the number of arguments expected, the cases for each argument, and an indication of which argument(s) the verb agrees with. This information is encoded as a string. (1) gives the valence information for *bhend* 'loosen', indicating that this verb takes two arguments. The most agent-like argument ("A") is marked with ergative case, the other (patient-like, "P") with nominative, and the verb will be inflected to agree with both of them.

(1) `\val A-ERG P-NOM V-a(A).o(P)`

[2] `http://www.spw.uzh.ch/clrp`.

This information reflects rich linguistic knowledge, the product of the analysis done by the annotators, and it is digitized. However, it is not really machine interpretable. While Toolbox can assist with morphological parsing (and thus with glossing of sentences), it does not make use of such syntactic information for syntactic parsing.

2.3 Precision Grammars for Language Documentation

Precision grammars are machine-readable sets of rules developed by hand to capture linguistic generalizations. Large-scale precision grammar projects have been carried out in a variety of linguistic frameworks, including HPSG (Pollard and Sag, 1994; Flickinger, 2000) (described further below), LFG (Kaplan and Bresnan, 1982; Butt et al., 2002)) and TAG (Joshi et al., 1975). DELPH-IN-style[3] HPSG grammars map surface strings to semantic representations in the format of Minimal Recursion Semantics (MRS; Copestake et al., 2005), and are reversible, i.e., suitable for use in both analysis (strings-to-MRS) and generation (MRS-to-strings).

Precision grammars can be deployed in transfer-based machine translation (e.g., Lønning et al., 2004), grammar checking applications (e.g., Suppes et al., 2012), and other NLP applications which benefit from a strong distinction between grammatical and ungrammatical strings (e.g., in generation) and/or detailed semantic representations. While broad coverage precision grammars can be expensive to build, the alternative of treebank-derived grammars presupposes resources which don't typically exist for endangered languages and are themselves costly to create. Furthermore, precision grammars, by locating analytical decisions in specific rules, can be more easily updated than treebanks, as more is understood about the language being described. Using the methodology of the Redwoods project (Oepen et al., 2004), precision grammars can be used to create treebanks which can be kept up to date with the grammar as it evolves. Both precision grammars and their associated treebanks can be valuable resources in language documentation (Bender et al., 2012).

2.4 The LinGO Grammar Matrix

As noted, precision grammars are time-consuming to develop. However, because similar structures recur across languages, the development time for new grammars can be reduced by repurposing grammar code developed for other languages. This is the idea behind multilingual grammar engineering projects, including the LinGO Grammar Matrix (Bender et al., 2002, 2010), ParGram (Butt et al., 2002; King et al., 2005), PAWS (Black and Black, 2009), and GF (Ranta, 2007). The Grammar Matrix stores a core grammar which includes (partial) analyses hypothesized to be cross-linguistically applicable, including basic phrase structure rule types for combining heads with different types of dependents, as well as an implementation of semantic compositionality, i.e., constraints which relate the semantic representation associated with a phrase to the semantic contributions of its daughters. In addition, the Grammar Matrix provides a series of libraries of analyses of cross-linguistically variable phenomena. These analyses are accessed through a web-based questionnaire which elicits a linguistic description of a language from a linguist and outputs a corresponding set of grammar files describing phrase structure rules, lexical rules, and lexical entries.

The information provided by the linguist is encoded in a plain text 'choices' file, where each 'choice' is a simple attribute-value pair. The customization system interprets the choices to output grammar files. The grammar files are encoded in the framework of Head-Driven Phrase

[3] `http://www.delph-in.net`

Structure Grammar (HPSG; Pollard and Sag, 1994), providing representations in the format of Minimal Recursion Semantics (Copestake et al., 2005) and are compatible with the DELPH-IN suite of grammar development and deployment tools, including the LKB (Copestake, 2002).

For the purposes of this work, the most important aspects of the Grammar Matrix are its support for the creation of lexical rules, which handle the ordering, basic form, and syntactico-semantic contributions of affixes, and its set of lexical types. Both the lexical rules and lexical types pair forms with complex feature structures. These feature structures encode syntactic and semantic information and are compatible with the feature structures for phrase structure rules, meaning that the lexical entries can be used as part of a grammar capable of both paring and generation.

The resources of the Grammar Matrix represent another rich source of linguistic knowledge, but in this case, the knowledge is not specific to a particular language. In order to create a grammar for a particular language, they need to be paired with information about that language: The forms and lexical meanings of individual words, their valence patterns, and the forms and effects of individual affixes. While affixes generally form a relatively small closed class, lexicons are another matter. The goal of this work is to see how effectively we can use the existing lexicographic work of the CLRP to flesh out a Grammar Matrix-derived lexicon for Chintang.

3 Methodology

3.1 Matrix Lexicons and Toolbox Lexicons

As described above, Toolbox lexicons are structured by user-designed fields (marked with initial tags) that store information including the orthography of a form, its gloss, example sentences, and any other information the lexicon developers would like to collect. In the case of the CLRP, this includes part of speech and detailed valence information (case and semantic roles). These are each encoded as a string in the value of the associated tag.

A lexicon for a DELPH-IN style grammar associates orthographic forms with complex feature structures representing morphological, syntactic and semantic information, encoded in such a way that this information can interact with lexical and phrase structure rules to license syntactic analyses of full sentences which furthermore embed compositionally created semantic representations. The relationship between the strings and these complex feature structures is mediated by lexical types which bear the constraints that describe the feature structures. The lexical types, in turn, are arranged into a multiple inheritance hierarchy so that each constraint need be stated only once and can be inherited by all lexical entries which require it.

```
\lex kond
\id 179
\psrev v
\val A-ERG P-NOM V-a(A).o(P)
\ge search; look.for
\dt 22/Feb/2011
```

Figure 1: Sample Toolbox entry from CLRP

A sample Toolbox entry is shown in Figure 1 while Figure 2 illustrates the corresponding Matrix entry. (Both are abbreviated, to focus on the most relevant information.) In Figure 1, the value of the tag `\lex` encodes the stem, `\ge` gives an English gloss, `\psrev` the part of speech, and `\val` the detailed valence information. Other fields not shown in the figure encode alternate forms of the entry, examples, and glosses in Nepali.

The Matrix entry in Figure 2, is a typed feature structure. The type of the whole structure is *trans-verb-lex*. This type provides (or inherits from its supertypes) most of the constraints on the entry. The only constraints provided directly in the lexical entry are the STEM value (*kond*, corresponding to `\lex` in Figure 1) and the PRED value *_search;look.for_v_rel*, i.e., the predicate symbol for the semantic relation associated from this entry. This is built on the basis of the `\ge` field of the Toolbox entry.

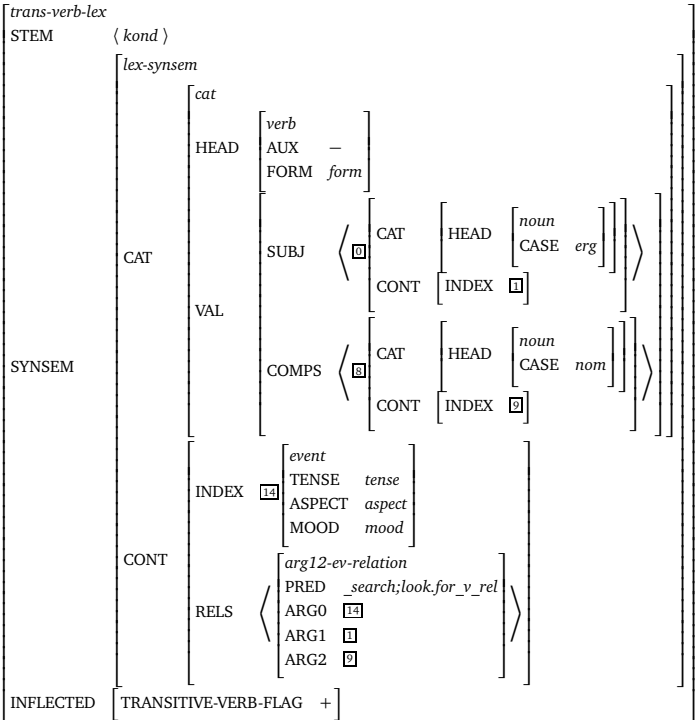

Figure 2: Sample Matrix entry corresponding to Figure 1

Turning to the information contributed by *trans-verb-lex*, the HEAD value indicates that this is a verb (and will head verbal projections such as VP and S), that it is not an auxiliary ([AUX −]), and that its form value is as yet underspecified ([FORM *form*]). When this lexical entry is inflected as either a finite verb or a non-finite verb, the FORM value will be constrained accordingly. The INDEX value is linked to the ARG0 of the relation contributed by the verb, and has underspecified values for TENSE, ASPECT, and MOOD. These, too, can be filled in via lexical rules for affixes that mark these values.

The VAL information indicates that this verb is seeking a subject and a complement, both headed

by nouns, where the subject must be in the ergative case and the object in the nominative.[4] Furthermore, the INDEX values of each are linked to the ARG1 and ARG2 positions in the semantic predicate, respectively. The number of arguments and the linking to the semantic roles is part of the cross-linguistic definition of a transitive verb in the Matrix. The constraints that both arguments are NPs (rather than, say, PPs) and the information about case come from specializations to the transitive verb type defined for Chintang by hand through the Grammar Matrix customization system questionnaire.

Finally, the feature INFLECTED is related to the morphotactic system. The value of this feature is a bundle of further 'flag' features (Goodman, 2012) tracking whether certain lexical rules have or have not applied, in order to encode dependencies between lexical rules and between lexical rules and lexical types. Here, we have shown only the TRANSITIVE-VERB-FLAG feature, whose + value will ensure that affixes throughout the affix chain will only be those that are compatible with transitive verbs.

3.2 Import of Lexical Entries from Toolbox Lexicons

The Grammar Matrix customization system provides facilities for the definition of lexical types. This is in principle unbounded: the user can define, for example, types for both common and proper nouns, as well as types for nouns of different genders and types for verbs with different case frames. The user can specify constraints on these types through the customization system (e.g., constraints on noun gender or case frames). Many other constraints, particularly those concerned with semantic composition, are inherited from the Matrix core grammar.

We extended the Grammar Matrix customization system to include a subpage that allows the user to define mappings between sets of properties encoded in a Toolbox lexical entry and user-defined lexical types. Figure 3 gives an example. This type maps entries from the Toolbox lexicon which are specified to have the part of speech 'v' and the valence 'S-NOM V-s(S)' to the type 'verb1'. This type is defined on another page of the customization system questionnaire to describe intransitive verbs with nominative case on their sole argument. Types inherited from the Matrix core grammar provide the constraints that contribute a one-place semantic predicate and link the sole syntactic argument to the semantic argument. The import facility creates the name symbol for that semantic predicate on the basis of the gloss or alternatively of the orthography of the stem (as specified by the user).

Since Toolbox allows users to define their own tags, the extension we designed for the Grammar Matrix customization system does not make any assumptions about the name or number of tags which will be relevant to each import class. Users fill in the name of the tag in the 'Toolbox tag' field for each tag-value pair, and can add arbitrarily many tag-value pairs with the 'Add' button. Another part of the page allows the user to specify the location of a Toolbox file to import from and upload it. Though Chintang was the initial test case for this import facility, there is (to our knowledge) nothing specific to Chintang nor the CLRP in the design of the system. It is available for use through the Grammar Matrix customization system's web-based

[4] The Grammar Matrix uses the names SUBJ and COMPS for the valence features, but makes relatively few assumptions about which properties accrue to the argument in SUBJ as opposed to those in COMPS cross-linguistically. For example, as case and agreement are both handled lexically, the system is flexible enough to model even tripartite case and agreement, where the sole argument of intransitives is handled differently from either argument of transitives (Drellishak, 2009). Similarly, grammar developers using the Grammar Matrix customization system can define multiple different classes within transitive and intransitive which behave differently with respect to agreement and/or case.

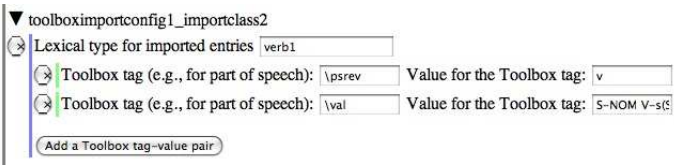

Figure 3: Sample Toolbox import class

questionnaire.[5]

The Grammar Matrix core grammar provides support for a wide variety of semantic valences. However, the present system only exposes simple intransitive and transitive valences to the customization page. As soon as the customization system is updated to expose more valence possibilities in the definition of lexical types, types using such valence possibilities will be available as targets for the import of corresponding Toolbox lexical entries, without any further updates to the extension we created. For the purposes of this study, however, we are limited to nouns and simple transitive and intransitive verbs.

The final 'choices' file for the Chintang grammar specifies import configurations for common nouns and two types each of transitive and intransitive verbs: native Chintang verbs which take the full range of inflectional morphology and verbs borrowed from Nepali which must co-occur with an auxiliary. This results in imported lexical entries for 4,741 common nouns, 282 native Chintang intransitive verbs, 142 borrowed Nepali intransitive verbs, 285 native Chintang transitive verbs, and 190 borrowed Nepali transitive verbs. Thus 5,640/9,034 (62%) of the entries in the Toolbox lexicon resulted in entries for the Matrix grammar, including 899/1,440 (62%) of the verbs. The most frequent remaining part of speech categories in the Toolbox lexicon file are adverbs (866), adjectives (515), interjections (377), and affixes (286).[6][7]

3.3 Implementation of Morphology

Chintang has a relatively complex morphological system, especially for verbal inflection. Schikowski (2012) identifies 12 suffix positions following verbs. On the other side of the stem, a verb root may take up to 4 prefixes, and additionally endoclitics can appear inside the prefix chain. One special type of prefix is that found in bipartite stems, where a specific prefix is idiosyncratically selected by the verb and is in effect part of the stem, though not always realized contiguously with the rest of the stem. To add to the complexity, the prefixes do not occur in a fixed order (Bickel et al., 2007). Beyond that, a single verbal word can contain up to four verb roots, each of which can host prefixes and suffixes. In these 'verb chains', any given prefix can appear only once, while suffixes are frequently repeated (Bickel et al., 2007; Schikowski, 2012). Finally, there is a host of morphophonological effects, some categorical

[5] `http://www.delph-in.net/matrix/customize/matrix.cgi`

[6] The Toolbox lexicon includes words from four different languages (Chintang, Nepali, Bantawa, and English), as words from all four of these languages appear in the collected data. These numbers reflect the full Toolbox file, and so are not directly representative of only Chintang. For example, Chintang has only two adjectives; the rest of the adjective entries come from other languages.

[7] These part of speech counts are based on the `\ps` field in the lexicon. The import of lexical entries was done based on the `\psrev` field. This field does not cover as many words as the older `\ps` field, but has been thoroughly reviewed.

and some variable, which change the surface form of any given morpheme depending on its phonological context (as well as on sociolinguistic factors).

The Grammar Matrix customization system provides extensive support for the definition of the morphotactic and morphosyntactic aspects of lexical rules, i.e., the order in which morphemes appear, co-occurrence restrictions between morphemes, and the syntactico-semantic constraints associated with each (Goodman, 2012). For this study, we defined a set of lexical rules through this system on the basis of the prose description of Chintang morphology in Schikowski 2012 and consultation between Schikowski (field linguist) and Bender (grammar engineer). Here we briefly describe the phenomena handled by this rule set.

There are a total of 160 verbal lexical rules (grouped into 54 position classes) and 24 nominal lexical rules (6 position classes) in the implementation. The position classes define the order of the affixes, including whether they are prefixes or suffixes and their relative order to other prefixes/suffixes. We handle verb chains by treating only the first verb root as an actual root in the model; this is facilitated by the fact that the first V position in a verb chain has the widest lexical variation. The current system allows up to three verb roots per chain, with the verb roots appearing in the second and third position treated as affixes. The position classes for each of these positions contain 32 rules, one for each verb root which can appear in non-initial position. The prefixes and suffixes which intervene between roots in the verb chain are treated as separate lexical rules for suffixes. This duplication of the lexical rule types is partially responsible for the high overall total of lexical rules.[8]

Elsewhere in the choices file we have specified information about the case system (number of cases and their names), possible values of tense, aspect and mood, possible values of person and number (including inclusive/exclusive distinctions in first person dual and plural), and other similar information. This allows us to model or partially model the syntactico-semantic effects of 131 of the 160 rules. The features targeted by these rules are shown in Table 1.[9] Examples of affixes whose syntactico-semantic effects are not modeled include the causative marker, possessive prefixes on nouns, and verb chain elements indicating the direction of motion or resulting position of a participant in the event. This information is not modeled because it is not directly supported by the customization system. The grammar output by the customization system is suitable for further hand-development, however, and nothing in HPSG theory or the DELPH-IN formalism would prevent encoding such information.

There are a few other ways in which this model of Chintang morphology is incomplete. First, it should be noted that we are abstracting away from most of the morphophonology by targeting the underlying representation given in the Toolbox files (both lexicon and corpus), rather than the transcription. Second, we are not modeling the phenomenon of free prefix order. This is possible in the DELPH-IN formalism but not supported by the Grammar Matrix customization system. Again, it would be relatively straightforward to modify the grammar to accommodate this, but we have chosen to focus our evaluation on the grammar as produced by the customization system. While the customization system can handle bipartite stems, and the facility for importing lexical entries from Toolbox anticipates this, we are not modeling them

[8] HPSG's type hierarchy in principle would allow us to define the morphosyntactic effects of these rules just once and cross-classify those types with the types encoding the position class information. The Grammar Matrix customization system interface, however, does not allow cross-classification of position class types with other kinds of types, so this kind of generalization must await hand-editing of the grammar files output by the customization system.

[9] Note that NEGATION isn't really a feature but rather a flag that causes the customization system to create a lexical rule which adds negation to the verb's semantic representation (Crowgey, ip).

Feature	# Rules	Notes
PERNUM	103	Person and/or number of verb's dependent or of noun
CASE	17	
FORM	35	Form (finite/non-finite) of verb
TENSE	27	
ASPECT	5	
MOOD	17	
NEGATION	8	

Table 1: Features constrained by lexical rules

at this time. Finally, while verb chains can in principle have four verbal roots, this model only allows up to three, because the four-root forms are extremely rare.

4 Evaluation

We created a choices file for Chintang which gives general grammatical information as well as definitions for lexical classes and lexical rules as described above. In addition, this choices file defines 11 closed-class lexical entries: 10 pronouns and one auxiliary. We used three sample narratives (totaling 2,906 word tokens) from the corpus as development data to refine the choices file. This process involved creating a grammar from the choices file using the customization system, loading the grammar into the LKB grammar development environment (Copestake, 2002) and processing the utterances in the narratives with the grammar using the [incr tsdb()] grammar profiling platform (Oepen, 2001). [incr tsdb()] provides facilities for browsing both results and errors encountered during parsing. These were used to identify forms that were not being handled appropriately. We then used the grammar exploration tools provided by the LKB to diagnose the source of the problem and then updated the choices file accordingly.

We then selected an additional four narratives to use as test data. The narratives range in length from 200 to 489 tokens (total: 1,453) and represent a range of domains: 'Durga_Exp' is a biographical monologue; 'pear_6-1' is a Pear Story (Chafe, 1980) elicited by asking the speaker to recount a story shown in a short, non-verbal film; 'story_rabbit' is a story about a clever rabbit who escaped a tiger; and 'choku_yakkheng' is a recipe for cooking nettle curry. We extracted the morpheme segmented line of each line in the narratives. An example is shown in (2), where the second line is the line we are targeting.

(2) thupro wassace uyuwakte pho
thupro wassak-ce u-yuŋ-a-yakt-e pho
many bird-NS 3NSS/A-live-PST-IPFV-IND.PST REP
'There lived many birds.' [ctn] story_rabbit.005

The performance of the grammar was evaluated in two ways. First, we evaluated coverage at both the type and token level over the test narratives. Table 2 gives the results. A word was counted as 'covered' if the grammar assigned it a morphological analysis that the grammar considered fully inflected. As shown in the table, the grammar found analyses for at least 50% of both word types and tokens across all the narratives, with the exception of 'story_rabbit' where the token-level coverage was only 35%. The ambiguity numbers in Table 2 reflect the

average over those words which had at least one analysis. These numbers reflect low ambiguity, with the maximal analyses per word form being only 8.

Narrative	total		# analyzed		% analyzed		avg ambiguity	
	type	token	type	token	type	token	type	token
Durga_Exp	206	489	120	265	58	54	1.24	1.14
choku_yakkheng	152	331	89	184	59	56	1.26	1.20
pear_6-1	206	433	105	203	56	51	1.20	1.62
story_rabbit	85	200	43	69	51	35	1.37	1.23
All	568	1453	324	721	57	50	1.40	1.27

Table 2: Coverage of customized grammar over test narratives

To get a sense of the accuracy of the resulting grammar, we randomly selected 10 word types from each of the four narratives (while ensuring that no word type was selected from more than one narrative). We used the LKB to parse each of these word types and compared the information in the resulting feature structure to the information in the gloss of the first instance of that word type in the narrative it was chosen from. We calculated precision and recall for each piece of grammatical information in the gloss and the feature structure.[10] In cases where the grammar found more than one analysis, we chose the best match to the gloss. The results are shown in Table 3.

Narrative	Total gold attributes	Precision	Recall	F-measure
Durga_Exp	16	.48	.88	.62
choku_yakkheng	21	.57	.62	.59
pear_6-1	31	.71	.92	.80
story_rabbit	29	.62	.83	.71
Total	97	.61	.82	.70

Table 3: Accuracy of customized grammar over 40 word types from test narratives

A large portion of the precision errors in this evaluation relate to cases where the grammar interprets the non-marking of some category as informative. For example, nouns that do not bear any affix for number as marked as singular in the grammar, and nouns not bearing any affix for case as nominative. These disagreements between the grammar and the glosses are counted as errors in Table 3, as the glosses are taken as the gold standard for this evaluation. However, the glosses reflect a systematic decision by the CLRP to not mark the contribution of zero morphemes. In most of these cases, the grammar is likely correct. Finally, verbs inflected for tense are considered to be finite by the grammar, and this is reflected in a (syntactic) feature FORM in addition to the semantic feature TENSE. The glosses mark non-finiteness explicitly, but do not mark finiteness separately from tense. Default singular number on nouns accounts for 19 errors, default nominative case 15, and finite form 9, of a total of 52 errors in precision. The 18 errors in recall are primarily due to cases where the intended lexical root is not available in the grammar but a homophone is.

Finally, we performed an error analysis to get a sense of the range of reasons a word form might

[10]For the gloss, we included all information provided. For the feature structure, we included only the predicate symbol from the root and any information that is added by some lexical rule in the grammar.

not be analyzed by the current grammar. We randomly selected 10 word forms from the four narratives that were not assigned analyses by the grammar. The failure of analysis of these 40 forms can be attributed to the following causes:

- [29 forms] Stems not imported to the grammar, because they don't match any of the import classes. These stems include verbs taking three arguments, adverbs, numerals, demonstratives, and other function words.
- [2 forms] Stems that are not in the version of the Toolbox lexicon used to import from.
- [4 forms] Affixes not implemented in the grammar.
- [5 forms] Other problems with the grammar, such as not allowing for case stacking and not allowing the affix order attested.

In general, we find the results of this evaluation encouraging: They suggest that the methodology presented here is effective at repurposing the results of the work on the Toolbox lexicon towards additional computational linguistic ends. Furthermore, the error analysis points the way towards effective means of improving the resulting grammar further, including fixing the specific errors with affixes that were identified, broadening the classes of verbs handled, adding adverbs, and creating lexical entries for high frequency closed-class words by hand.

5 Related Work

This work is similar in spirit to Bender's (2008) development of an implemented grammar for Wambaya (ISO639-3: wmb) based on the Grammar Matrix and a descriptive grammar. However, that work focused on hand-development of the grammar and included a manually entered lexicon, in contrast to our work on automatically populating the lexicon for the implemented grammar.

Other work applying grammar engineering and shared resources (including typological information) to endangered or other resource poor languages includes the Parser and Writer for Syntax system (PAWS; Black and Black, 2009) and Linguist's Assistant (Beale, 2011). We are not aware of any work addressing lexicon repurposing for these systems, but methodology analogous to what we propose in this paper should be applicable to them as well.

More generally, our work is situated within a broader context of reuse of lexical resources across formalisms and across systems. Other work along these lines includes the work of Kamei et al. (1997) and Bond et al. (2009) on making it possible to share user dictionaries across different MT systems, that of Bond et al. (2008) on repurposing a variety of resources (both WordNets and other lexical resources) in order to create a WordNet for Japanese and that of (McConville and Dzikovska, 2007) on creating lexical entries for a TRIPS grammar (Dzikovska, 2004) on the basis of FrameNet (Baker et al., 1998).

Conclusion and perspectives

The target lexical entries for a precision grammar derived from the Grammar Matrix are much more complex than the information explicitly encoded in even a thorough Toolbox lexicon. The work of developing the Toolbox lexicon is, however, the hard part. In this paper we have shown how it is possible to use a language-independent (i.e., explicitly multi-lingual) tool to leverage

the effort and linguistic analysis encoded in a Toolbox lexicon to create the kind of resource required for a machine-readable, precision grammar.

However, it is important to note that this is only a first step. Previous work building medium to large scale grammars with the DELPH-IN technology, including the broad-coverage English Resource Grammar (Flickinger, 2000, 2011) and a medium-sized grammar for Wambaya (Bender, 2008) suggest that it should indeed be possible to build a substantial grammar fragment for Chintang that uses this lexicon. The Wambaya grammar is especially pertinent for two reasons: first, like Chintang, it represents an application of the Grammar Matrix to a language not considered in its initial development, and second, its lexical types are based on the same general supertypes as those developed here for Chintang. Nonetheless, every language is different, and it is not possible to know without building the grammar whether the lexical types will be compatible with the specific grammatical phenomena attested in Chintang. We intend to develop such a grammar to test this in future work.

Acknowledgments

CPDP was funded by the VW Foundation as part of the DoBeS program (Grant No. II/79092 to Bickel) and CLRP by the VW Foundation (Grant No. II/81 730 to S. Stoll), the German Research Foundation (Grants No. BI 799/5-1 and BI 799/9-1 to B. Bickel) and the Swiss National Science Foundation (Grant Nr. 100012_140881 to Bickel).

In addition, this material is also based upon work supported by the National Science Foundation under Grants No. 0644097 and 1160274 to Bender. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We are grateful to David Wax for assistance in integrating the software for importing lexical entries from Toolbox to the Grammar Matrix into the Grammar Matrix's web-based questionnaire.

References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Beale, S. (2011). Using linguist's assistant for language description and translation. In *Proceedings of the IJCNLP 2011 System Demonstrations*, pages 5–8, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Bender, E. M. (2008). Evaluating a crosslinguistic grammar resource: A case study of Wambaya. In *Proceedings of ACL-08: HLT*, pages 977–985, Columbus, Ohio. Association for Computational Linguistics.

Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L., and Saleem, S. (2010). Grammar customization. *Research on Language & Computation*, pages 1–50. 10.1007/s11168-010-9070-1.

Bender, E. M., Flickinger, D., and Oepen, S. (2002). The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In Carroll, J., Oostdijk, N., and Sutcliffe, R., editors, *Proceedings of the Workshop*

on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics, pages 8–14, Taipei, Taiwan.

Bender, E. M., Ghodke, S., Baldwin, T., and Dridan, R. (2012). From database to treebank: Enhancing hypertext grammars with grammar engineering and treebank search. In Nordhoff, S., editor, *Electronic Grammaticography* (In press). University of Hawaii Press, Hawaii.

Bickel, B., Banjade, G., Gaenszle, M., Lieven, E., Paudyal, N., Rai, I., Rai, M., Rai, N. K., and Stoll, S. (2007). Free prefix ordering in Chintang. *Language*, 83(1):43–73.

Bickel, B., Gaenszle, M., Rai, N. K., Lieven, E., Banjade, G., Bhatta, T. N., Paudyal, N., Pettigrew, J., Rai, I. P., Rai, M., Schikowski, R., and Stoll, S. (2009). Audiovisual corpus of the chintang language, including a longitudinal corpus of language acquisition by six children, plus a trilingual dictionary, paradigm sets, grammar sketches, ethnographic descriptions, and photographs. *DOBES Archive*, `http://www.mpi.nl/DOBES`.

Black, C. A. and Black, H. A. (2009). PAWS: Parser and writer for syntax: Drafting syntactic grammars in the third wave. In *SIL Forum for Language Fieldwork*, volume 2.

Bond, F., Isahara, H., Kanzaki, K., and Uchimoto, K. (2008). Boot-strapping a WordNet using multiple existing WordNets. In *Proceedings of LREC 2008*, pages 1619–1624.

Bond, F., Okura, S., Yamamoto, Y., Murata, T., Uchimoto, K., Kato, M., Shimazu, M., and Suzuki, T. (2009). Sharing user dictionaries across multiple systems with UTX-S. In *Second International Workshop on Intercultural Collaboration (IWIC-2009)*, Stanford CA.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The parallel grammar project. In Carroll, J., Oostdijk, N., and Sutcliffe, R., editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 1–7.

Chafe, W. (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex Pub. Corp., Norwood NJ.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, CA.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(2–3):281–332.

Crowgey, J. (ip). The syntactic exponence of sentential negation: A model for the LinGO grammar matrix. Master's thesis, University of Washington.

Drellishak, S. (2009). *Widespread But Not Universal: Improving the Typological Coverage of the Grammar Matrix*. PhD thesis, University of Washington.

Dzikovska, M. O. (2004). *A Practical Semantic Representation for Natural Language Parsing*. PhD thesis, University of Rochester, Rochester NY.

Ebert, K. (2003). Kiranti languages: an overview. In Thurgood, G. and LaPolla, R., editors, *The Sino-Tibetan languages*, chapter 31, pages 505–517. Routledge, London/New York.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6 (1) (Special Issue on Efficient Processing with HPSG):15 – 28.

Flickinger, D. (2011). Accuracy v. robustness in grammar engineering. In Bender, E. M. and Arnold, J. E., editors, *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 31–50. CSLI Publications, Stanford, CA.

Goodman, M. W. (2012). Generation of machine-readable morphological rules from human-readable input. Unpublished ms, University of Washington.

Joshi, A., Levy, L. S., and Takahashi, M. (1975). Tree adjunct grammars. *Journal of Computer Systems Science*.

Kamei, S., Itoh, E., Fujii, M., Hirai, T., Saitoh, Y., Takahashi, M., Hiyama, T., and Muraki, K. (1997). Shareable formats and their supporting environments for exchanging user dictionaries among different MT systems as part of AAMT activities. In *MT Summit VI*.

Kaplan, R. M. and Bresnan, J. (1982). Lexical-Functional Grammar: a formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*. MIT Press.

King, T. H., Forst, M., Kuhn, J., and Butt, M. (2005). The feature space in parallel grammar writing. *Research on Language and Computation, Special Issue on Shared Representations in Multilingual Grammar Engineering*, 3(2):139–163.

Lønning, J. T., Oepen, S., Beermann, D., Hellan, L., Carroll, J., Dyvik, H., Flickinger, D., Johannessen, J. B., Meurer, P., Nordgård, T., Rosén, V., and Velldal, E. (2004). LOGON. A Norwegian MT effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*.

McConville, M. and Dzikovska, M. O. (2007). Extracting a verb lexicon for deep parsing from FrameNet. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 112–119, Prague, Czech Republic. Association for Computational Linguistics.

Michailovsky, B. (1994). Manner vs. place articulation in the kiranti initial stops. In Kitamura, H., Nishida, T., and Nagano, Y., editors, *Current issues in Sino-Tibetan linguistics*, pages 766–772. National Museum of Ethnology, Osaka.

Oepen, S. (2001). [incr tsdb()] — Competence and performance laboratory. User manual. Technical report, Saarland University, Saarbrücken, Germany.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. *Journal of Research on Language and Computation*, 2(4):575–596.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press and CSLI Publications.

Rāī, N. K., Rāī, M., Paudyāl, N. P., Schikowski, R., Bickel, B., Stoll, S., Gaenszle, M., Banjade, G., Rāī, I. P., Bhaṭṭa, T. N., Sauppe, S., Rāī, R. M., Rāī, J. K., Rāī, L. K., Rāī, D. B., Rāī, G., Rāī, D., Rāī, D. K., Rāī, A., Rāī, C. K., Rāī, Ś. M., Rāī, R. K., Pettigrew, J., and Dirksmeyer, T. (2011). *Chintāṅa-Nepālī-Ãgrejī śabdakośa tathā vyākaraṇa*. Chintang Language Research Programme, Kāṭhmāḍauṁ.

Ranta, A. (2007). Modular grammar engineering in GF. *Research on Language & Computation*, 5:133–158.

Schikowski, R. (2012). Chintang morphology. Unpublished ms, University of Zürich.

Stoll, S., Bickel, B., Lieven, E., Banjade, G., Bhatta, T. N., Gaenszle, M., Paudyal, N. P., Pettigrew, J., Rai, I. P., Rai, M., and Rai, N. K. (2012). Nouns and verbs in chintang: children's usage and surrounding adult speech. *Journal of Child Language*, 39:284 – 321.

Suppes, P., Flickinger, D., Macken, B., Cook, J., and Liang, T. (2012). Description of the EPGY Stanford University online courses for mathematics and language arts. In *International Society for Technology in Education Annual (ISTE) 2012 Conference*, San Diego CA.